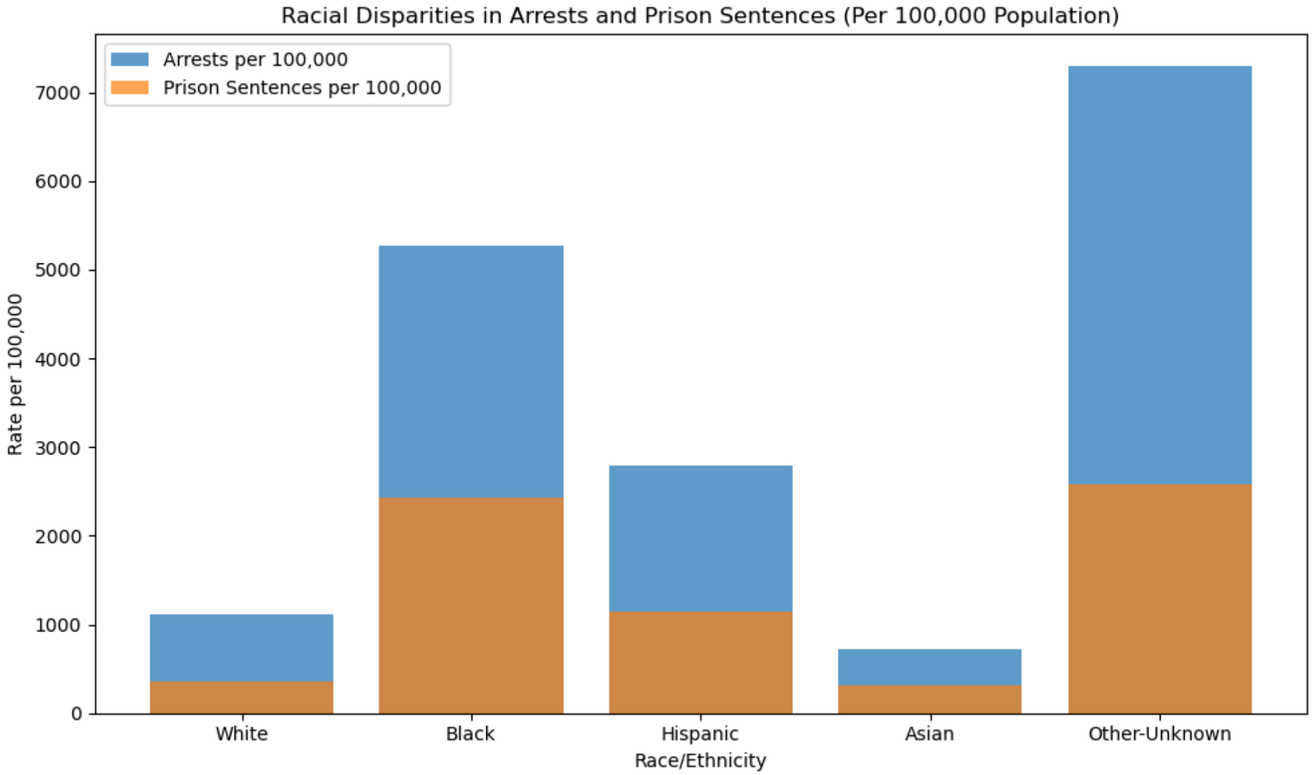
PROGRESS REPORT

TABLE OF CONTENTS

INTRODUCTION

Current Analysis of the situation



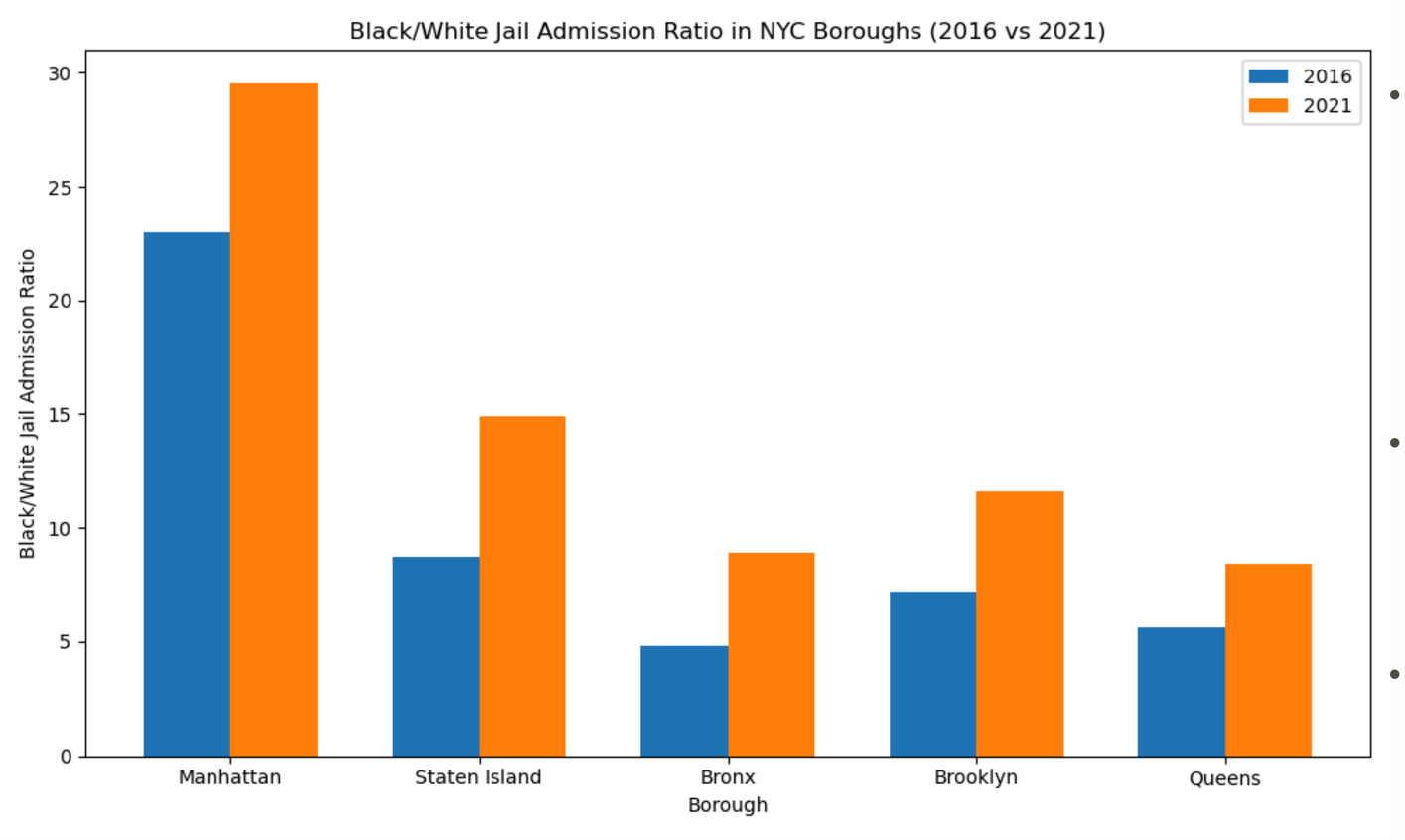
1. New York State Division of Criminal Justice Services. (2023). 2022 Population, arrests, prison by race. NYS DCJS-OJRP.
2. Monaghan, S., Rempel, M., & Lin, T. (2023, February). Racial disparities in the use of jail across New York City, 2016-2021. Data Collaborative for Justice.
3. McCormack, S., & Barber, J. (n.d.). A racial disparity across New York that is truly jarring.

Summary:

- Black individuals have the highest rate of both arrests and prison sentences per 100,000 people, with prison sentences significantly higher compared to other groups.
- Hispanic individuals also experience a notable disparity, with prison sentence rates higher than the White and Asian populations.
- White individuals make up the majority of the population but have significantly lower arrest and prison sentence rates.

INTRODUCTION

Current Analysis of the situation



1. New York State Division of Criminal Justice Services. (2023). 2022 Population, arrests, prison by race. NYS DCJS-OJRP.
2. Monaghan, S., Rempel, M., & Lin, T. (2023, February). Racial disparities in the use of jail across New York City, 2016-2021. Data Collaborative for Justice.
3. McCormack, S., & Barber, J. (n.d.). A racial disparity across New York that is truly jarring.

Summary:

- **Higher Arrest Rates in Specific Boroughs:** Black individuals face disproportionately high rates of incarceration in all New York City boroughs, with these rates increasing over time. Brooklyn and Staten Island saw the largest increases in Black jail admissions between 2016 and 2021. In 2021, Brooklyn had the highest number of admissions, while Staten Island had the fewest. However, Black individuals were still more likely to be incarcerated in all boroughs compared to White individuals.

- **Neighborhood-Level Disparities:** High-poverty areas, disproportionately populated by Black and Brown residents, contribute significantly to the city's jail population. Black men in these neighborhoods experience higher rates of repeated incarcerations compared to other racial groups.

- **Bail and Pretrial Disparities:** Geographic differences also affect pretrial incarceration, with Black individuals disproportionately unable to afford bail. Black individuals represent 58% of those admitted to jail due to an inability to pay bail, further exacerbating racial and geographic inequalities.

INTRODUCTION

Business Context

Our project focuses on addressing the significant issue of bias within the U.S. criminal justice system through the use of advanced data analytics and machine learning techniques.

Given the vast amount of data generated in law enforcement—from arrest records to demographic information—there is a an opportunity to leverage this data to uncover trends, especially related to systemic biases.

The project specifically focus on how factors like race, age, gender, geography, and political climates (e.g., liberal vs. conservative regions) influence arrest patterns. By identifying these trends, the project aims to provide data-driven insights that can guide reforms in law enforcement and judicial practices, thereby helping reduce racial, geographic, and socio-economic biases in arrests.

The broader business context lies in the societal need for fairness in law enforcement, as public scrutiny over racial and demographic disparities in arrests has increased especially in the US.

Therefore, this project fits into a larger agenda of promoting fairness, equity, and data transparency in US the judicial system.

With the proper application of predictive models, the project aims to mitigate the biases that currently exist and provide recommendations for better, more informed policy-making.

INTRODUCTION

Target Audience

- **Policy-makers and Government Officials:** those individuals are directly responsible for enacting laws and making decisions on resource allocation in the criminal justice system. They could use the findings of this project to drive legislative changes that promote fairer treatment across different demographics and regions.

- **Law Enforcement Agencies:** Police departments, sheriffs, and other local law enforcement agencies could use the predictive models to re-evaluate arrest procedures and potentially adjust their training or resource allocation to minimize biased practices.

- **Criminal Justice Reform Advocates:** Non-profit organizations and advocacy groups such as the ACLU or The Sentencing Project that work on criminal justice reform would be key stakeholders. They could use this data to highlight systemic problems and push for change within the judicial system.

- **Data Scientists and Legal Researchers:** Academic and research institutions focused on criminal justice data, machine learning, and fairness would benefit from the methodologies developed in this project and could apply them to further research on criminal justice reform.

INTRODUCTION

Potential Stakeholders

- **Federal, State, and Local Government Bodies:** Agencies that manage law enforcement and policy creation are the key stakeholders, as they can directly implement changes based on the project’s findings.
- **Community Leaders:** Those representing populations disproportionately affected by biased arrests, such as minority and marginalized communities, would be essential stakeholders in ensuring that the findings lead to actionable reforms.
- **Law Enforcement Organizations:** Including police departments and unions that could see shifts in practices based on the predictive model's results.
- **Public Interest and Advocacy Groups**: Organizations like the NAACP or the ACLU, who work to address disparities in arrest patterns and push for legal reforms, would find these insights valuable for their advocacy campaigns.
- **Technology and Data Vendors**: If the model proves successful, vendors offering tools to manage and implement predictive analytics in law enforcement agencies may also be stakeholders.

INTRODUCTION

Aim of our project

Our primary goal for this project is to create a predictive model that helps estimate the likelihood of an individual being arrested for various criminal offenses, taking into account factors like demographics, geography, and political environments. By using arrest data, public records, and external demographic information, we aim to identify trends and uncover patterns of bias in the system. With these insights, our goal is to provide recommendations that can help reduce inequities and promote fairness within the criminal justice system.

We also want this project to be part of a bigger conversation about fairness in machine learning. Beyond just revealing arrest trends, we'll be testing fairness metrics like Disparate Impact Ratio, Equal Opportunity Difference, and Demographic Parity to ensure that the model doesn't reinforce existing biases. In the end, we hope to offer actionable recommendations to help shape policies that lead to more just and equitable law enforcement practices.

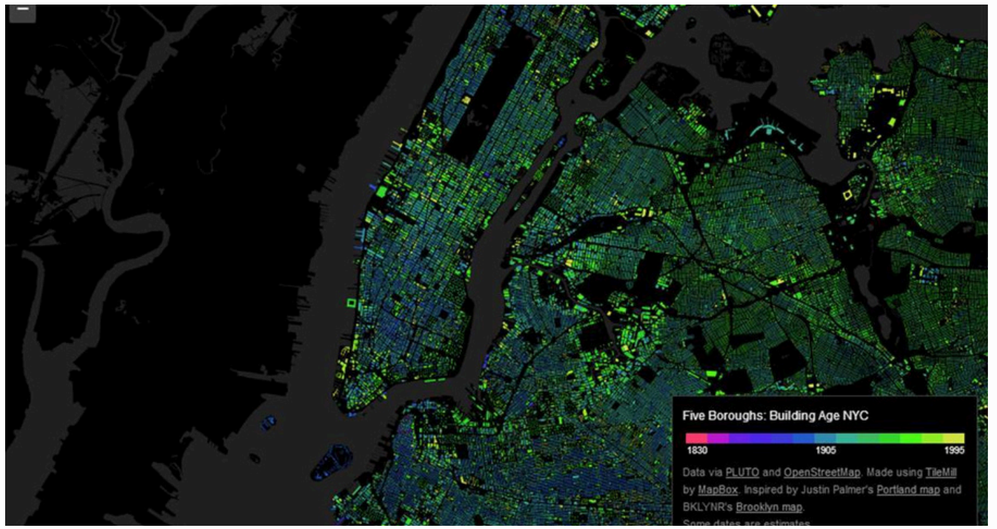
DESCRIPTION OF THE DATASET

01 **Stop, Question and Frisk Data.**
Last Updated: year 2023
Data records from the NYPD Stop.

https://www.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page

02 **Neighborhood Financial Health Digital Mapping and Data.**
Last Updated: June 27, 2022
Data Provided By the Department of Consumer and Worker Protection (DCWP).

https://data.cityofnewyork.us/Business/Neighborhood-Financial-Health-Digital-Mapping-and-/r3dx-pew9/about_data



1- STOP, QUESTION AND FRISK DATA.

In the NYPD dataset for the year 2023, each row in the data represents a unique stop-and-frisk incident recorded by the NYPD. The columns provide detailed information about various aspects of these incidents.

Incident Details:

Date, time, location (precinct, sector, address, etc.), duration, reason for the stop (radio run, self-initiated, etc.)

Officer and Supervisor Information:

Rank, command code, and whether the supervising officer reviewed the activity log entry.

Suspect Demographics:

Age, sex, race, height, weight, build, eye color, hair color, and other identifying descriptions.

Circumstances of the Stop:

Suspected crime, presence of weapons, actions of the suspect (casing, concealed possession, etc.), background circumstances (violent crime, suspect known to carry a weapon, etc.).

Justification and Basis for the Stop:

Whether the officer explained the stop, the suspect's demeanor, the basis for any searches conducted (admission, consent, hard object, etc.).

Outcomes of the Stop:

Whether the suspect was arrested, frisked, searched, if a summons was issued, and if any contraband or weapons were found.

Use of Force:

Whether any physical force was used during the stop, including CEW, firearm, handcuffs, OC spray, etc.

2- NEIGHBORHOOD FINANCIAL HEALTH DIGITAL MAPPING AND DATA.

This provide an overview of financial health within neighborhoods across New York City. Each row in the dataset represents a specific Public Use Microdata Area (PUMA) in NYC, which is a statistical geographic unit with at least 100,000 people. The columns provide insights into various aspects of financial well-being in these areas.

Geographic and Demographic Information:

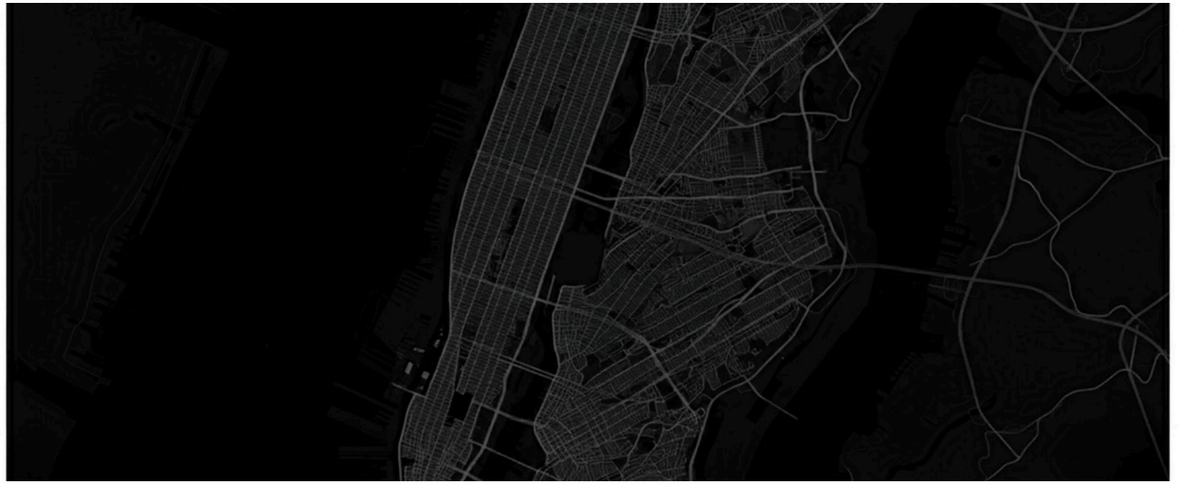
Borough and Neighborhood details. Community Districts within each PUMA. Poverty rate, median income, and racial/ethnic composition of the population.

Five key financial health Indicators:

- Access to affordable and high-quality financial services.
- Access to affordable and essential goods and services.
- Access to quality jobs and income supports.
- Stability of housing and ability to manage financial shocks.
- Opportunities to build assets and plan for the future.

Specific indicators for each Indicators:

- Homeownership rates.
- Access to banks and credit unions.
- Job training and placement support.
- Housing affordability.
- Health insurance coverage.
- Retirement security.
- Community resources.



Performance and Ranking:

- Overall index score for each financial health goal.
- Ranking of each PUMA compared to others based on the index score.
- Outcome or raw score for each indicator.
- Ranking of each PUMA based on its performance on individual indicators.

INITIAL DATA EXPLORATION

```
nypd_2023_filtered = nypd_2023.loc[:, [
    'STOP_ID',
    'STOP_FRISK_DATE',
    'STOP_FRISK_TIME',
    'DAY2',
    'STOP_WAS_INITIATED',
    'ISSUING_OFFICER_COMMAND_CODE',
    'ISSUING_OFFICER_RANK',
    'SUPERVISING_OFFICER_RANK',
    'SUPERVISING_OFFICER_COMMAND_CODE',
    'SUSPECTED_CRIME_DESCRIPTION',
    'STOP_DURATION_MINUTES',
    'OBSERVED_DURATION_MINUTES',
    'OFFICER_EXPLAINED_STOP_FLAG',
    'OFFICER_NOT_EXPLAINED_STOP_DESCRIPTION',
    'SUSPECT_ARRESTED_FLAG',
    'SUSPECT_ARREST_OFFENSE',
    'DEMEANOR_OF_PERSON_STOPPED',
    'SUSPECT_REPORTED_AGE',
    'SUSPECT_SEX',
    'SUSPECT_RACE_DESCRIPTION',
    'SUSPECT_HEIGHT',
    'SUSPECT_WEIGHT',
    'SUSPECT_BODY_BUILD_TYPE',
    'SUSPECT_EYE_COLOR',
    'SUSPECT_HAIR_COLOR',
    'STOP_LOCATION_SECTOR_CODE',
    'WEAPON_FOUND_FLAG'
]]

nypd_2023_filtered.replace('(null)', np.nan, inplace=True)
```

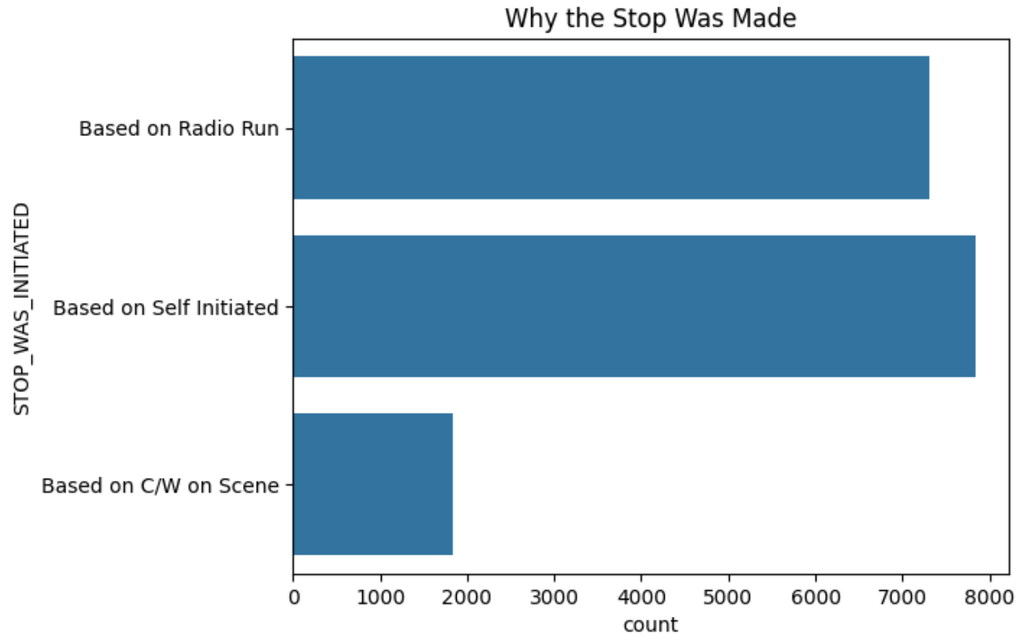
Initial Data Wrangling

- Use 2023 NYC Stop and Frisk data as starting sample and will scale up
- Select relevant variables for analysis
- Standardize values and handle missing data (NAs) for future analysis

Link to Google Colab File

https://colab.research.google.com/drive/12y0BEcxbK7qU41H7r2nuFhDv4nwyk1_1?usp=sharing

INITIAL DATA EXPLORATION

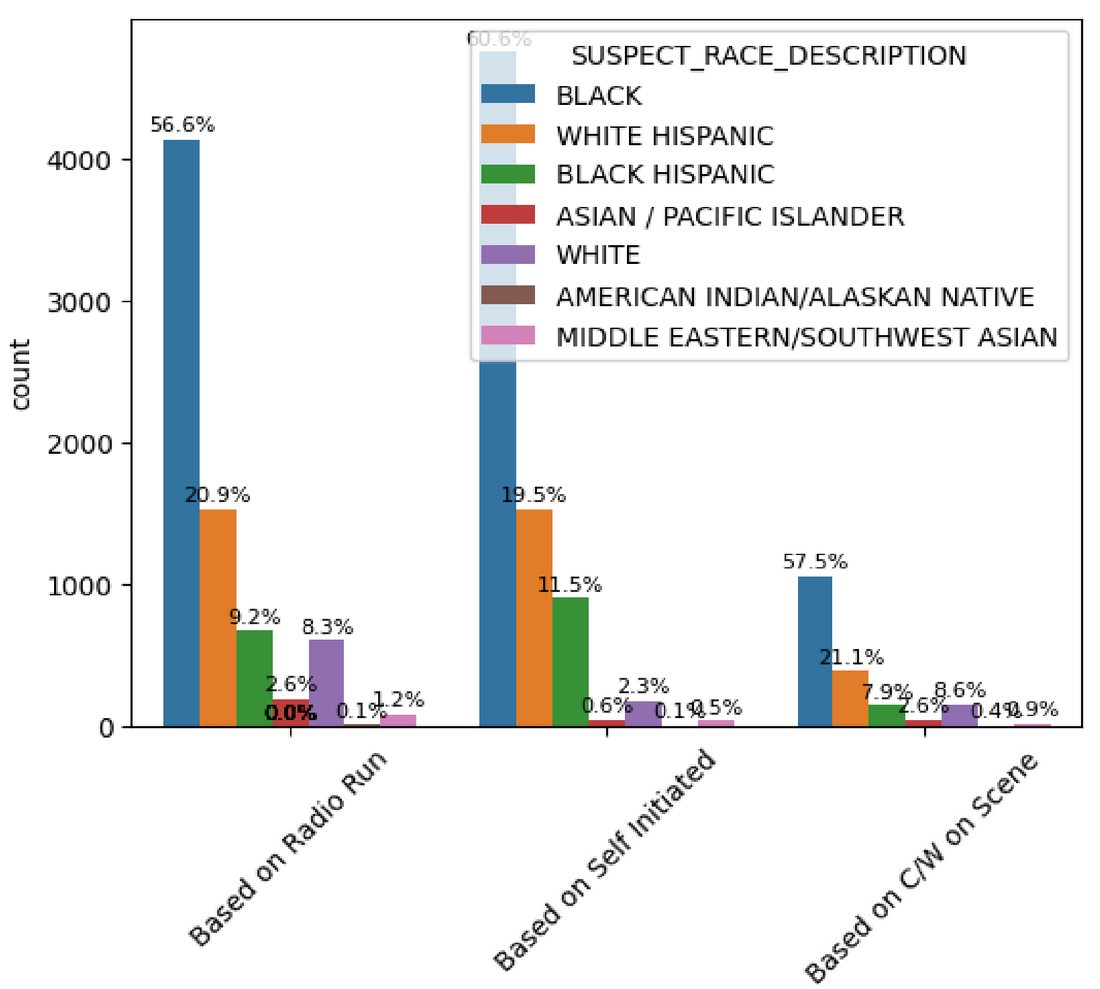


01 Based on the plot, most stops in 2023 New York City (nearly 8,000) were independently initiated by officers, while about 7,000 resulted from dispatch or radio calls.

02 A list of the most common suspected crimes shows that Criminal Possession of a Weapon (CPW) is the most frequent offense in the 2023 data. Further analysis on crime types and weapon possession will follow.

SUSPECTED_CRIME_DESCRIPTION	
CPW	8978
ROBBERY	1703
ASSAULT	1296
PETIT LARCENY	1090
BURGLARY	869
GRAND LARCENY AUTO	561

INITIAL DATA EXPLORATION



Based on the 2023 U.S. Census, white individuals make up 37.5% of New York City's population, while African Americans constitute 23.1%.

However, for all three reasons for being stopped, African Americans represent over 56% of those stopped, with the proportion increasing for stops initiated independently by officers. In contrast, white individuals are stopped at a much lower rate compared to their demographic proportion, with only 2.3% of stops being self-initiated by the police.

This discrepancy is also evident when comparing white and Black Hispanic individuals across all three categories of stops.

Lhttps://www.census.gov/quickfacts/newyorkcitynewyork

INITIAL DATA EXPLORATION

SUSPECT_ARRESTED_FLAG	count
N	12071
Y	4900

01 Out of 16,971 stop incidents in 2023, 4,900 suspects were arrested, accounting for approximately 29% of the total, based on NYC data.

02 According to the table, 60% of those arrested in 2023 NYC stop-and-frisk data are Black. This is largely due to the high number of Black individuals being stopped, as their arrest rate is only 29%, compared to 38% for white individuals, who make up just 7% of arrests. This suggests that Black individuals are more likely to be stopped without having committed a crime, highlighting potential biases in the stopping process.

Racial Group Percentages Among Arrested and Non-Arrested Individuals

SUSPECT_RACE_DESCRIPTION SUSPECT_ARRESTED_FLAG	AMERICAN INDIAN/ALASKAN NATIVE	ASIAN / PACIFIC ISLANDER	BLACK	BLACK HISPANIC	MIDDLE EASTERN/SOUTHWEST ASIAN	WHITE	WHITE HISPANIC
N	0.0	0.02	0.61	0.11	0.01	0.05	0.21
Y	0.0	0.02	0.60	0.09	0.01	0.07	0.21

Arrest and Non-Arrest Percentages by Racial Group

SUSPECT_RACE_DESCRIPTION SUSPECT_ARRESTED_FLAG	AMERICAN INDIAN/ALASKAN NATIVE	ASIAN / PACIFIC ISLANDER	BLACK	BLACK HISPANIC	MIDDLE EASTERN/SOUTHWEST ASIAN	WHITE	WHITE HISPANIC
N	0.64	0.62	0.71	0.76	0.72	0.62	0.7
Y	0.36	0.38	0.29	0.24	0.28	0.38	0.3

INITIAL DATA EXPLORATION

SUSPECTED_CRIME_VS_ARREST_OFFENSE	count
Y	3956
N	944

01 Of the 4,900 arrested suspects, 3,956 (80.7%) had suspected crimes that matched their arrest offenses.

The next steps involve analyzing the 944 cases where crimes did not match the arrest offenses, examining racial dynamics, and identifying potential biases.

02 Among those suspected of Criminal Possession of a Weapon (CPW), 22% were found carrying a weapon, while 8% of individuals stopped without CPW suspicion were also found carrying a weapon.

WEAPON_FOUND_FLAG	CPW N	CPW Y
N	0.92	0.78
Y	0.08	0.22

INITIAL DATA EXPLORATION

OBSERVED DURATION TIME

SUSPECT_RACE_DESCRIPTION	count	mean	std	min	25%	50%	75%	max
AMERICAN INDIAN/ALASKAN NATIVE	22.0	1.0	1.2	0.0	0.0	1.0	1.0	5.0
ASIAN / PACIFIC ISLANDER	288.0	1.7	2.7	0.0	1.0	1.0	2.0	20.0
BLACK	9939.0	4.9	234.3	0.0	1.0	1.0	2.0	23042.0
BLACK HISPANIC	1719.0	1.8	10.9	0.0	1.0	1.0	2.0	440.0
MIDDLE EASTERN/SOUTHWEST ASIAN	142.0	1.1	1.3	0.0	0.0	1.0	1.0	10.0
WHITE	942.0	3.3	47.7	0.0	1.0	1.0	2.0	1452.0
WHITE HISPANIC	3447.0	1.5	2.6	0.0	1.0	1.0	2.0	55.0

The table shows that Black individuals had the longest time under officer observation before being stopped, followed by white individuals. Native Americans and individuals of Middle Eastern and Southeast Asian descent had the shortest observed times.

INITIAL DATA EXPLORATION

STOP DURATION TIME

SUSPECT_RACE_DESCRIPTION	count	mean	std	min	25%	50%	75%	max
AMERICAN INDIAN/ALASKAN NATIVE	22.0	12.9	10.9	1.0	5.0	10.5	15.0	50.0
ASIAN / PACIFIC ISLANDER	288.0	10.8	10.5	1.0	5.0	9.0	15.0	70.0
BLACK	9939.0	7.8	12.7	0.0	2.0	5.0	10.0	581.0
BLACK HISPANIC	1719.0	8.0	15.5	0.0	2.0	5.0	10.0	339.0
MIDDLE EASTERN/SOUTHWEST ASIAN	142.0	13.5	12.2	1.0	5.0	10.0	20.0	63.0
WHITE	942.0	10.9	13.7	0.0	5.0	8.0	13.0	183.0
WHITE HISPANIC	3447.0	8.9	14.8	0.0	2.0	5.0	10.0	275.0

The table shows that Black individuals and Black Hispanics had the shortest stop durations, which is notable given their low arrest rates, potentially reflecting systemic racism in the practice of stopping them. Meanwhile, Native Americans and individuals of Middle Eastern and Southeast Asian descent experienced the longest stop durations, which warrants further investigation.

NEXT STEPS

EXPLORATORY DATA ANALYSIS (EDA)- NEIGHBORHOOD FINANCIAL HEALTH DATA

01

The next phase begins with additional wrangling of the Neighborhood Financial Health Data. This involves addressing any missing values and creating new features, such as composite indices for financial vulnerability and employment opportunities. These variables will help enrich the analysis when combined with the stop-and-frisk data.

In parallel, we will perform EDA on the neighborhood dataset by calculating descriptive statistics and visualizing distributions of key indicators like poverty rate, median income, and access to services. Geospatial mapping of financial health will also help identify neighborhood-level patterns, laying the groundwork for further analysis. This step ensures that the neighborhood data is well-understood before merging with the stop-and-frisk data.

NEXT STEPS
COMBINING AND ANALYZING DATASETS

02 After preparing the datasets individually, the next step is to merge them using geographic identifiers. This combined dataset will allow us to analyze the relationship between financial health and police behavior. We will then perform additional EDA on this merged data, exploring correlations between poverty levels, racial demographics, and stop-and-frisk rates.

Feature selection will focus on identifying the most impactful variables, such as racial composition, neighborhood financial scores, and stop outcomes.

NEXT STEPS
MODELING AND HYPERPARAMETER TUNING

03

Once data is combined and features selected, we will build predictive models to estimate arrest likelihood. We will explore models such as Logistic Regression for its interpretability, Random Forest, and Gradient Boosting for capturing complex relationships. In selecting the final model, we will balance interpretability and accuracy. We may lean more towards interpretability, as it will help us understand the key drivers behind arrests and biases, which is crucial for actionable insights and policy recommendations.

We will use GridSearch for hyperparameter fine-tuning to optimize model performance. Regarding performance metrics, our primary focus will be on balancing recall (the ability to correctly identify actual arrests) and precision (minimizing false positives), as both are critical in this context. The F1-score will provide a balanced metric that combines both precision and recall, which is ideal for evaluating the overall effectiveness of the model.

Thank you